Learning from Data, Tutorial Sheet for week 10

School of Informatics, University of Edinburgh

Instructor: Amos Storkey

1. Suppose a hypothetical UK railservice from Edinburgh to Oldfort is often subject to delays. The train service is run by three different train operating companies (TOC). Over the course of a year, a random sample of the services was taken. The following data was obtained

	Weather	Season	TOC	Day	Lateness
Case 1	Windy	Summer	RotRail	Weekday	On time
Case 2	Windy	Winter	GNAF	Weekday	Delayed
Case 3	Windy	Autumn	GNAF	Weekday	Delayed
Case 4	Calm	Summer	Virgo	Weekend	Delayed
Case 5	Windy	Winter	RotRail	Weekend	Delayed
Case 6	Calm	Summer	Virgo	Weekday	Delayed
Case 7	Calm	Spring	RotRail	Weekday	On time
Case 8	Windy	Autumn	GNAF	Weekend	Delayed
Case 9	Calm	Winter	Virgo	Weekend	Delayed
Case 10	Calm	Spring	Virgo	Weekday	Delayed
Case 11	Windy	Autumn	GNAF	Weekday	Delayed
Case 12	Windy	Spring	GNAF	Weekday	On time
Case 13	Windy	Summer	RotRail	Weekday	On time
Case 14	Calm	Autumn	RotRail	Weekday	On time
Case 15	Windy	Winter	RotRail	Weekday	Delayed
Case 16	Calm	Autumn	Virgo	Weekday	Delayed
Case 17	Windy	Summer	Virgo	Weekday	Delayed
Case 18	Windy	Spring	Virgo	Weekend	Delayed
Case 19	Calm	Winter	GNAF	Weekday	On time
Case 20	Calm	Spring	GNAF	Weekend	On time

Find the root (top) node selected using the maximum information gain tree building procedure to classify whether a train will be delayed or on time. Show that it selects according to which TOC is providing the service.

You might find the following table a helpful starter

	Delayed	On time
Calm	5	4
Windy	8	3
Summer	3	2
Winter	4	1
Autumn	4	1
Spring	2	3

	Delayed	On time
RotRail	2	4
GNAF	4	3
Virgo	7	0
Weekday	8	6
Weekend	5	1

The maximum information gain tree building procedure creates the following first two layers of the tree. Suppose the whole tree were pruned to this level (2 layers). Find the final decision tree by filling in the missing classification values and missing classification ratios below

```
                             TOC
              |---------------^-------------|-----------------|
            RotRail                        GNAF              Virgo
              |                             |                 |
            Season                        Season           Delayed
   |------|---^--|------|         |------|---^--|------|
Summer Winter Autumn Spring    Summer Winter Autumn Spring
  |       |      |      |         |      |      |      |
On time Delay                     ?      ?    Delay
        (2:0)  (0:1)  (0:1)    (0:0)  (1:1)
```

2. Using your decision tree from question 1, how would you classify

	Weather	Season	TOC	Day	Lateness
Example 1	Windy	Autumn	RotRail	Weekday	?
Example 2	Calm	Summer	Virgo	Weekday	?
Example 3	Calm	Spring	GNAF	Weekend	?

3. If the attribute 'day' were to be replaced by a different attribute, such as 'date', what do you think the maximum information gain tree building procedure would do? Is there a problem with this? What does this say about the maximum information tree building method?

Comment on your results from question 1 in the light of this.